Three Types of Average Dependency Distances of Sentences in a Multilingual Parallel Corpus

Masanori Oya
Meiji University
Masanori_oya2019@meiji.ac.jp

Abstract

This paper is an exploration of three types of dependency distances of the sentences in a multilingual parallel corpus, in order to verify the expectation that we can obtain an objective, quantitative measure to indicate cross-linguistic variation of the syntactic-structural setting of human languages. The results indicated that pair-wise average dependency distances seem to categorize languages into several groups, and type-wise average dependency distances seem to provide us with fine-grained quantification of syntactic properties of individual natural languages.

1. Introduction

Dependency distance is the linear distance between two words in dependency relationship within the same sentence. It has been investigated as one of the important measures of memory burden and syntactic complexity (Gibson, 1998, 2000; Gildea and Temperley, 2010; Grodner and Gibson, 2005; Liu 2007, 2008; Liu et al., 2017; Oya, 2013). For example, in the sentence "David read three articles yesterday," the noun David depends on read as the subject, and the dependency distance between them is 1, and the noun articles depends on read as the object, and the dependency distance between them is 2 (the direction of dependency is ignored here, and dependency distances are given as absolute values). In the example sentence above, the dependency distance between the subject and the verb is shorter than that between the object and the verb.

Investigations of dependency distances have been conducted from the viewpoint of finding out the general properties of human languages in general, and it has been pointed out that natural languages show a certain preference for shorter dependency distances; for example, Gibson (2000) proposes the Dependency Locality Theory, which basically states that the preference for shorter dependency distances in natural languages is due to the limit of short-term memory during the integration of words into a larger structure. Liu (2008) shows that, based on the corpus-based investigation of 20 languages, there is a threshold of dependency distances which is 4, and the average dependency distance of these natural languages is shorter than that of an artificial language in which the dependency relationships of words are randomized. Futrell et al. (2015) also reported the similar result based on the corpus data of 37 natural languages.

Provided that shorter dependency distances are preferred across different languages and they are below a certain threshold which has been found in previous research, we can assume that it is worth searching for cross-linguistic variations of dependency distances, not only by focusing on the dependency distances in a language as a whole, but also by focusing on the difference of average dependency distances of language pairs, and also on the dependency distances of different dependency types, such as the dependency distance between the object and the verb which is longer than that between the subject and the verb, as shown in the example above. By doing this, it is

expected that we can obtain an objective, quantitative measure to indicate cross-linguistic variation of the syntactic-structural setting of human languages.

This paper reports an attempt to verify this expectation by exploring the dependency distances of the sentences in a multilingual parallel corpus. The structure of this article is as follows: Section 2 introduces the idea of average dependency distance of a sentence. Section 3 further describes the three types of average dependency distances. Section 4 and 5 describe the data and procedure used in the study, respectively. Section 6 reports the results, which are discussed in Section 7. Section 8 concludes this article.

2. The average dependency distance of a sentence as a quantitative measure of structural difference between sentences

This section introduces the idea of the average dependency distance (henceforth ADD) of a sentence, and proposes to use it as a quantitative measure for structural difference between sentences. The ADD of a sentence equals the sum of all the dependency distances in the sentence divided by the number of the dependencies in the same sentence. The ADDs of two sentences in a translation pair of two languages indicate the structural and quantitative difference of these sentences. If the ADD of one of the sentences in a translation pair is shorter than the ADD of the other in the same translation pair, it means that shorter dependency distances are used in the former than the latter when expressing the same meaning, as far as this translation pair is concerned. If we have more than one translation pair of two languages, say Language A and Language B, and the number of the translation pairs of Language A and B is large enough, we may find that the ADD of A is shorter than that of B, and we may conclude from this fact that shorter dependency distances tend to be used in A than B when expressing the same meaning.

With this in mind, not only can one translation pair be used to calculate the ADDs of the sentences in the pair, a large number of sentences in a parallel corpus of a variety of languages can be the data for calculating the ADDs of the sentences in the corpus. A parallel corpus contains sentences with the same meaning across different languages. For example, from a parallel corpus of 20 languages with 1,000 sentences, we can obtain 1,000 sentence groups, and each of these sentence groups contains sentences from 20 languages which share the same meaning. We can consider the ADDs calculated from such data as the quantitative measure of the structural difference of these languages which is controlled in terms of their semantics. This measure is obviously more reliable than that obtained from the ADDs of sentence pairs taken randomly from these two languages.

3. Three types of ADD

This section introduces the three different ways to calculate ADDs of the sentences in a parallel corpus. These ADDs are aligned from the most generic one (the ADD of a particular language in general) to a more specific one (the ADD of a particular dependency type of a language).

First, the ADD can be calculated from the whole parallel corpus, as explained above; The ADD of sentences in a corpus of a language equals the sum of all the dependency distances in the sentences in the corpus divided by the number of the dependencies in the corpus. Then, the ADDs of the languages in a parallel corpus quantitatively represent the structural differences among them as a whole. I propose to call the ADD thus calculated the language-wise ADD (henceforth LADD) of the language.

Second, the ADDs of two languages in a parallel corpus can also be calculated by the sum of the difference between ADDs of each translation pair of these two languages divided by the number of these translation pairs. I propose to call the ADD thus calculated from a parallel corpus the pair-wise ADD (henceforth PADD). The PADD of two languages in a parallel corpus contains more information than the LADDs of the two languages in the same parallel corpus, because the PADD takes into consideration the semantic parallelism between the sentences in each translation pair of these two languages.

The PADD of a pair of Languages A and B can be either more than or less than zero; it is more than zero if the ADD of Language A is longer than that of Language B, and less than zero if the ADD of Language A is shorter than that of Language B. A longer PADD of Language A with respect to

Language B means that Language A tends to contain longer dependency distances than Language B when expressing the same meaning.

The absolute value of the PADD of a language pair can be used to show the similarity of these two languages; If it is shown that the absolute value of the PADD of Language A with respect to Language B is smaller than the absolute value of the PADD of Language A with respect to Language C, then it can be interpreted that Language A is closer to Language B than Language C in terms of their PADDs.

Third, along with the LADD and PADD, we can also calculate the average dependency distance of each of the dependency types, such as the dependency distance between a verb and its subject or object, between the verb of a main clause and the verb of its subordinate clause, or between a noun and the verb of a relative clause which modifies the noun. I propose to call them type-wise ADD (henceforth TADD); it is expected that different languages show longer or shorter TADDs of the same dependency types. For example, the TADD between a verb and its subject must be longer in SOV languages than that in SVO languages, because there is a higher possibility of the existence of an object between a verb and its subject in SOV languages, resulting in longer TADD between a verb and its subject in SOV languages, compared to that in SVO languages in which there is little (or zero) possibility of the existence of an object between a verb and its subject. TADDs of different dependency types across different languages will provide us with quantitative measures for their more fine-grained structural differences which have been difficult, if not impossible, to detect before. Notice that it is impossible to calculate the pair-wise TADD of two languages, because one dependency type used in a sentence of Language A can be absent from its translation pair of Language B; for example, the subject of the main clause in Language A can be translated as a zero subject in Language B. In such cases, it is impossible to calculate the difference of the ADDs of this dependency type across this translation pair of Language A and B.

4. Data

The data used in this paper is the annotated sentences in Parallel Universal Dependencies Treebanks 2.7 (henceforth PUD). These treebanks have been created for the purpose of shared task on Multilingual Parsing from Raw Text to Universal Dependencies at CoNLL 2017 (http://universaldependencies.org/conll17/). PUD contains 20,000 sentences from 20 different languages (Arabic, Chinese, Czech, English, Finnish, French, German, Hindi, Indonesian, Icelandic, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Thai, and Turkish), and the number is growing as new languages are included. The subcorpus of PUD for each language contains 1,000 sentences, in a fixed order across languages, and aligned basically one-to-one across languages, except for some sentences which are translated into more than one sentence. Thus, PUD provides us with aligned translation pairs across 20 languages. PUD ultimately contains 361,000 translation pairs; in PUD, there are 1,000 sentences for each language. Each of these 1,000 sentences in one language is aligned to its translation counterparts in 19 other languages. Thus, one language in PUD has 19,000 translation pairs. This calculation applies to all the other 19 languages in PUD; therefore, PUD contains 361,000 translation pairs (19,000 multiplied by 19). Of the 1,000 sentences, 750 are translated from English texts, and the remaining 250 sentences are translated from German, French, Italian, or Spanish, which had been translated into English, and then further translated into other languages (their ID numbers indicate the original language). The translation of these sentences has been conducted by professional translators, and annotated with morphological and syntactic tags by Google. Then, UD community members convert the annotation into Universal Dependencies, according to the UD Ver. 2 guidelines. For further details on PUD treebanks, refer to the UD webpage (https://universaldependencies.org/).

5. Procedure

We can obtain the dependency distance of each dependency relation in PUD through the dependency tag on every word in a sentence. It is conducted by an original Ruby script, then the output of each language is used to calculate for the word count and the ADD of each sentence. This result is employed to calculate the LADD of the language. This process is conducted for every language in PUD, and then the result is employed to calculate the PADD of two language pair. In this

study, the LADDs of the 20 languages in PUD are calculated. Then, the PADDs are calculated for all the possible language pairs of the 20 languages in PUD. Lastly, the TADDs of seven languages are calculated (these languages are chosen due to the results of PADDs; explained in Section 6.3) with respect to the following dependency types: (1) between a verb and its subject; (2) between a verb and its object; (3) between a noun and a phrase modifying the noun; (4) between the verb of a main clause and the verb of a subordinate clause; and (5) between a noun and the verb of a relative clause which modifies the noun. These dependency types are focused on in this study for the following reasons: The first two dependency types are chosen here because they are typical dependency types between a verb and a noun (phrase), the third one is chosen because it is one of the most frequent dependency types, and the last two types are chosen because they represent the embeddedness of sentences.

6. Results

6.1 The LADDs of the 20 languages in PUD

First, the descriptive statistics of the dependency distances the sentences of all 20 languages in PUD are shown in Table 1. These languages are aligned from that with the largest LADD to the smallest. It should be pointed out that, all across the 20 languages in PUD, their mode of dependency distances is 1, and the majority of their median are 2. The LADDs of Chinese, Japanese, and Korean fall within the interval between 3.5 and 4, and their SDs are within the five largest, along with those of Hindi and German; the LADDs of Romance languages in PUD (French, Italian, Portuguese, and Spanish) fall within the interval between 3.4 and 3.5; the LADDs of Slavic languages in PUD (Czech, Polish, and Russian) fall within the interval between 3.2 and 3.4. The LADDs of Germanic languages in PUD (English, German, Icelandic, and Swedish) do not fall within an interval as narrow as Romance and Slavic languages.

A Kruskal-Wallis test was performed and the result shows that there are statistically significant differences among the LADDs of these languages ($H(19)= 3665.18$, $p < 0.1$).

The results of multiple comparisons by a Steel-Dwass test show that, out of the all 190 pairs, significant differences are found between 149 language pairs, while no significant differences are found between the following 41 pairs, indicating that languages of these pairs are closer to each other as far as their LADDs are concerned; Arabic and Indonesian, Arabic and Icelandic, Arabic and Polish, Arabic and Russian, Czech and Spanish, Czech and Korean, Czech and Russian, Czech and Swedish, German and Hindi, English and Spanish, English and French, English and Italian, English and Korean, English and Portuguese, English and Turkish, Spanish and French, Spanish and Italian, Spanish and Korean, Spanish and Portuguese, Spanish and Turkish, Finnish and Indonesian, Finnish and Icelandic, Finnish and Polish, French and Italian, French and Korean, French and Portuguese, French and Turkish, Indonesian and Icelandic, Indonesian and Polish, Indonesian and Russian, Icelandic and Polish, Icelandic and Russian, Italian and Korean, Italian and Portuguese, Italian and Turkish, Japanese and Chinese, Korean and Portuguese, Korean and Turkish, Polish and Russian, Portuguese and Turkish, Russian and Swedish.

	Dependencies	Sum of DD	LADD	Median	Mode	Min.	Max.	SD
Hindi	23829	100142	4.2025	2	1	1	53	5.522
German	21329	88375	4.1434	2	1	1	51	4.884
Chinese	21415	85815	4.0072	2	1	1	47	5.013
Japanese	28784	110599	3.8424	2	1	1	72	6.446
Turkish	16882	59798	3.5421	1	1	1	40	4.661
English	21168	74452	3.5172	2	1	1	56	4.278
Korean	16584	58137	3.5056	1	1	1	46	4.923
French	24727	85996	3.4778	2	1	1	52	4.756
Italian	23731	81791	3.4466	2	1	1	63	4.644
Portuguese	23388	80406	3.4379	2	1	1	61	4.589
Spanish	23280	79580	3.4184	2	1	1	59	4.575
Czech	18603	63088	3.3913	2	1	1	45	4.080
Swedish	19071	63431	3.326	2	1	1	50	4.082
Russian	19355	63467	3.2791	2	1	1	46	4.118
Arabic	20751	66968	3.2272	1	1	1	51	4.642
Polish	18389	59106	3.2142	2	1	1	42	4.022
Indonesian	19440	61731	3.1755	2	1	1	46	4.063
Finnish	15813	49885	3.1547	2	1	1	44	3.495
Icelandic	18828	59088	3.1383	2	1	1	49	3.853
Thai	22322	60134	2.6939	1	1	1	42	3.279

Table 1. Descriptive statistics of the average dependency distances of the 20 languages in PUD

Figure 1 is the box plots of the ADDs of the 20 languages in PUD. These languages are aligned from the largest LADD to the left and the smallest LADD to the right. This indicates that languages with larger LADDs such as German, Hindi, and Chinese also show a wide variation of average dependency distances, while the variation of average dependency distances is small for languages with smaller LADDs such as Icelandic and Thai.

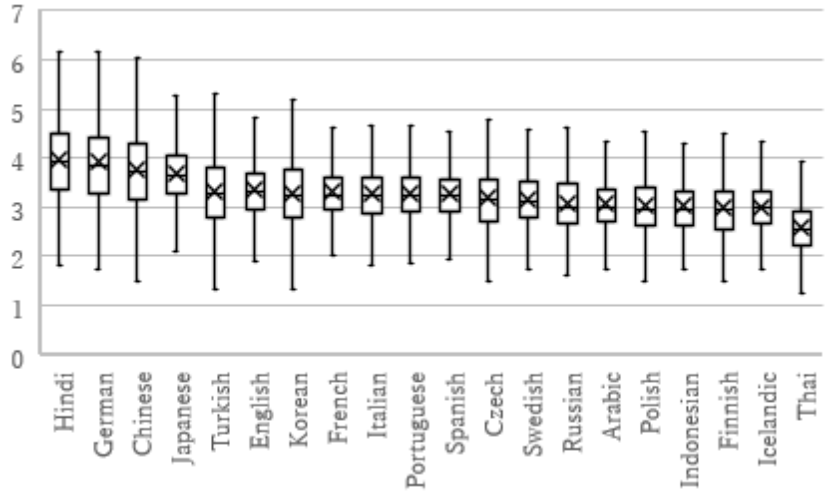

Figure 1. Distributions of ADDs of the 20 languages in PUD

6.2 The PADDs of the 20 languages in PUD

The PADDs of all the possible pairs of the 20 languages in PUD are shown in Table 2. The PADD of Language A and Language B is calculated by subtracting the ADD of the sentence of Language B from that of its translation pair of Language A, then the ADDs of the 1,000 sentences pairs of language A and B are summed and then divided by 1,000, which is the number of translation pairs. In Table 2, PADDs are boldface when the language pairs are found non-significant according to the results of multiple comparisons by the Steel-Dwass test mentioned in the previous section, indicating that the languages of those pairs are relatively closer to each other in terms of their PADD.

	Arabic	Chinese	Czech	English	Finnish	French	German	Hindi	Indonesian	Icelandic	Italian	Japanese	Korean	Polish	Portuguese	Russian	Spanish	Swedish	Thai	Turkish
Arabic	n/a	-0.70	-0.13	-0.29	0.09	-0.25	-0.84	-0.90	**0.04**	**0.06**	-0.21	-0.62	-0.23	**0.02**	-0.20	**-0.02**	-0.19	-0.10	0.47	-0.26
Chinese	0.70	n/a	0.57	0.40	0.79	0.45	-0.15	-0.20	0.74	0.76	0.49	**0.08**	0.47	0.72	0.50	0.68	0.51	0.60	1.17	0.44
Czech	0.13	-0.57	n/a	-0.16	0.22	-0.12	-0.71	-0.77	0.17	0.19	-0.08	-0.49	**-0.10**	0.15	-0.07	**0.11**	**-0.06**	**0.03**	0.60	-0.13
English	0.29	-0.40	0.16	n/a	0.38	**0.05**	-0.55	-0.60	0.34	0.36	**0.08**	-0.32	**0.06**	0.32	**0.09**	0.27	**0.10**	0.20	0.76	**0.03**
Finnish	-0.09	-0.79	-0.22	-0.38	n/a	-0.34	-0.93	-0.98	**-0.04**	**-0.02**	-0.30	-0.70	-0.32	**-0.06**	-0.29	-0.11	-0.28	-0.19	0.38	-0.35
French	0.25	-0.45	0.12	**-0.05**	0.34	n/a	-0.60	-0.65	0.29	0.31	**0.04**	-0.37	**0.02**	0.27	**0.05**	0.23	**0.06**	0.15	0.72	**-0.01**
German	0.84	0.15	0.71	0.55	0.93	0.60	n/a	**-0.05**	0.89	0.91	0.63	0.23	0.61	0.87	0.64	0.82	0.65	0.75	1.31	0.58
Hindi	0.90	0.20	0.77	0.60	0.98	0.65	**0.05**	n/a	0.94	0.96	0.68	0.28	0.67	0.92	0.69	0.87	0.70	0.80	1.37	0.63
Indonesian	**-0.04**	-0.74	-0.17	-0.34	**0.04**	-0.29	-0.89	-0.94	n/a	**0.02**	-0.26	-0.66	-0.27	**-0.02**	-0.25	**-0.07**	-0.24	-0.14	0.43	-0.31
Icelandic	**-0.06**	-0.76	-0.19	-0.36	**0.02**	-0.31	-0.91	-0.96	**-0.02**	n/a	-0.28	-0.68	-0.29	**-0.04**	-0.27	**-0.09**	-0.26	-0.16	0.41	-0.33
Italian	0.21	-0.49	0.08	**-0.08**	0.30	**-0.04**	-0.63	-0.68	0.26	0.28	n/a	-0.40	**-0.02**	0.24	**0.01**	0.19	**0.02**	0.12	0.68	**-0.05**
Japanese	0.62	**-0.08**	0.49	0.32	0.70	0.37	-0.23	-0.28	0.66	0.68	0.40	n/a	0.39	0.64	0.41	0.59	0.42	0.52	1.09	0.35
Korean	0.23	-0.47	**0.10**	**-0.06**	0.32	**-0.02**	-0.61	-0.67	0.27	0.29	**0.02**	-0.39	n/a	0.25	**0.03**	0.21	**0.04**	0.13	0.70	**-0.03**
Polish	**-0.02**	-0.72	-0.15	-0.32	**0.06**	-0.27	-0.87	-0.92	**0.02**	**0.04**	-0.24	-0.64	-0.25	n/a	-0.23	**-0.05**	-0.22	-0.12	0.45	-0.29
Portuguese	**0.20**	-0.50	0.07	**-0.09**	0.29	**-0.05**	-0.64	-0.69	0.25	0.27	**-0.01**	-0.41	**-0.03**	0.23	n/a	0.18	**0.01**	0.11	0.67	**-0.06**
Russian	**0.02**	-0.68	**-0.11**	-0.27	0.11	-0.23	-0.82	-0.87	**0.07**	**0.09**	-0.19	-0.59	-0.21	**0.05**	-0.18	n/a	-0.17	**-0.07**	0.49	-0.24
Spanish	0.19	-0.51	**0.06**	**-0.10**	0.28	**-0.06**	-0.65	-0.70	0.24	0.26	**-0.02**	-0.42	**-0.04**	0.22	**-0.01**	0.17	n/a	0.10	0.66	**-0.07**
Swedish	0.10	-0.60	**-0.03**	-0.20	0.19	-0.15	-0.75	-0.80	0.14	0.16	-0.12	-0.52	-0.13	0.12	-0.11	**0.07**	-0.10	n/a	0.57	-0.17
Thai	-0.47	-1.17	-0.60	-0.76	-0.38	-0.72	-1.31	-1.37	-0.43	-0.41	-0.68	-1.09	-0.70	-0.45	-0.67	-0.49	-0.66	-0.57	n/a	-0.73
Turkish	0.26	-0.44	0.13	**-0.03**	0.35	**0.01**	-0.58	-0.63	0.31	0.33	**0.05**	-0.35	**0.03**	0.29	0.06	0.24	**0.07**	0.17	0.73	n/a

Table 2. The PADDs of the language pairs from PUD

Table 2 shows that the absolute values of almost all the boldface PADDs are less than 0.1 (only one exception is the pair of Czech and Russian). Recall that these pairs have been found that the difference of their LADDs is not significantly different, as the result of a Steel-Dwass test. Though the PADD of a language pair is not calculated in the same way as the multiple comparisons by a Steel-Dwass test, the results seem to be similar with each other. This seems to suggest that a small absolute value of the PADD between two languages reflects statistically significant similarity between them.

The PADDs in Table 2 also indicate that the 20 languages in PUD are divided into four groups in terms of the closeness to each other represented by the small absolute value of their PADD; the first group includes German and Hindi, the second includes Chinese and Japanese, the third includes English, Korean, Romance languages (French, Italian, Portuguese, and Spanish), and Turkish, and the fourth includes Arabic, Finnish, Indonesian, Icelandic, Polish, Russian, and Swedish. Czech seems to belong to the second and the third group, because it is close to two of the second-group languages (Korean and Spanish), and two of the third-group languages (Russian and Swedish). Thai does not belong to any of these groups, because the absolute values of its PADDs are all larger than 0.1.

6.3 The TADDs of five dependency types of seven languages in the PUD

This section shows the TADDs of the five dependency types (subjects, objects, noun-modifying phrases, adverbial clauses, and relative clauses) of seven languages (Chinese, English, German, Hindi, Japanese, Korean, and Thai). These languages are chosen here according to the groups of PADDs mentioned in Section 6.2; two languages from the first group (German and Hindi), two languages from the second group (Chinese and Japanese), two languages from the third group (English and Korean), and Thai.

First, Table 3 shows the descriptive statistics of the dependency distances between verbs and their subjects (abbreviated as NSUBJs, following the convention of UD) of these languages, including their TADDs, and it indicates that Japanese has the longest TADD for NSUBJ (9.995) among these

seven languages, and its SD is also the largest (9.147), while English has the shortest TADD (3.055) and the smallest SD (2.956):

	NSUBJs	Sum of DD	TADD	Median	Mode	Min.	Max.	SD
Chinese	1846	7926	4.294	2	1	1	30	4.428
English	1631	4982	3.055	2	1	1	29	2.956
German	1688	7658	4.537	3	1	1	31	4.232
Hindi	1300	9872	7.594	6	1	1	46	6.044
Japanese	1484	14833	9.995	7	2	2	49	9.147
Korean	1706	10209	5.984	4	1	1	42	5.798
Thai	1691	5566	3.292	2	1	1	27	3.246

Table 3. The descriptive statistics of the dependency distances between verbs and their subjects in Chinese, English, German, Hindi, Japanese, Korean, and Thai

Second, Table 4 shows the descriptive statistics of the dependency distances between verbs and their objects (abbreviated as OBJs), and it indicates that German has the longest TADD (3.619) and Hindi the largest SD (3.020), while Thai has the shortest TADD (1.254) and English the smallest SD (1.149).

	OBJs	Sum of DD	TADD	Median	Mode	Min.	Max.	SD
Chinese	1526	5170	3.388	3	1	1	29	2.824
English	876	1939	2.213	2	2	1	9	1.149
German	895	3239	3.619	3	1	1	17	2.955
Hindi	1457	3576	2.454	1	1	1	29	3.020
Japanese	839	2362	2.815	2	2	2	36	2.581
Korean	1030	1881	1.826	1	1	1	26	2.167
Thai	1734	2174	1.254	1	1	1	30	1.363

Table 4. The descriptive statistics of the dependency distances between verbs and their objects in Chinese, English, German, Hindi, Japanese, Korean, and Thai

Third, Table 5 shows the descriptive statistics of the dependency distances between nouns and the phrases which modify these nouns, and it indicates that the TADDs of Thai is the longest (3.484) and Chinese has the largest SD (2.495), while Korean has the shortest TADD (1.766) and German the smallest SD (1.453).

	NMODs	Sum of DD	TADD	Median	Mode	Min.	Max.	SD
Chinese	702	2254	3.211	2	2	1	37	2.495
English	1498	4382	2.925	3	2	1	25	1.733
German	1373	3491	2.543	2	2	1	15	1.453
Hindi	1540	4181	2.715	2	2	1	21	2.130
Japanese	2287	7155	3.129	2	2	1	31	2.260
Korean	655	1157	1.766	1	1	1	16	1.460
Thai	945	3292	3.484	3	2	2	21	2.097

Table 5. The descriptive statistics of the dependency distances between nouns and the phrases modifying these nouns in Chinese, English, German, Hindi, Japanese, Korean, and Thai

Next, Table 6 shows the descriptive statistics of the dependency distances between the verbs of main clauses and the verbs of adverbial clauses modifying these main clauses (abbreviated as Advcls), and it indicates that German and Hindi have the longest TADDs (10.255 and 10.056, respectively) and Japanese has the largest SD (8.308), while Korean has the shortest TADD (5.587) and the smallest SD (5.13).

	Advcls	Sum of DD	TADD	Median	Mode	Min.	Max.	SD
Chinese	516	3637	7.048	5.5	2	1	41	5.545
English	293	2345	8.003	7	8	1	28	4.787
German	220	2256	10.255	9	2	1	29	6.373
Hindi	198	1991	10.056	9	8	2	27	5.653
Japanese	903	8542	9.460	8	2	1	59	8.308
Korean	998	5576	5.587	4	1	1	35	5.13
Thai	321	2937	9.150	8	6	2	32	5.578

Table 6. The descriptive statistics of the dependency distances between the verbs of main clauses and adverbial clauses modifying these main clauses in Chinese, English, German, Hindi, Japanese, Korean, and Thai

Lastly, Table 7 shows the descriptive statistics of the dependency distances between nouns and the verbs of relative clauses which modify these nouns (abbreviated as Relcls), and it indicates that Hindi has the longest TADD (12.967) and German has the second longest TADD (9.524), while Korean has the shortest TADD (1.745) and the smallest SD (1.379).

	Relcls	Sum of DD	TADD	Median	Mode	Min.	Max.	SD
Chinese	448	1859	4.150	3	1	2	26	2.626
English	1119	4118	3.680	4	2	1	18	3.129
German	271	2581	9.524	8	8	3	36	4.729
Hindi	215	2788	12.967	12	9	3	37	6.026
Japanese	211	1008	4.777	3	2	1	40	3.916
Korean	1188	2073	1.745	1	1	1	14	1.379
Thai	613	2556	4.170	3	2	1	26	6.74

Table 7. The descriptive statistics of the dependency distances between nouns and relative clauses modifying these nouns in Chinese, English, German, Hindi, Japanese, Korean, and Thai

7. Discussion

7.1 Preference for shorter dependency distances and possible variations of LADDs according to language families

The modes and medians of the dependency distances of all across the 20 languages in PUD clearly support the claim made by the previous research cited above (e.g., Futrell et al. 2015) that human languages prefer shorter dependency distances. In addition to this, the LADDs of these languages do not exceed 4 (except for German and Hindi), which is the threshold of dependency distances across 20 languages shown in Liu (2008). Along with this preference for shorter dependency distances across languages, we can notice the fact that languages of the same language family may share similar LADDs which fall within a narrow

interval (e.g., Romance languages), but not always (e.g., Germanic languages). This can be either a coincidence caused by some unexpected bias in the corpus data, or manifestation of certain cross-linguistic property which has yet to be revealed by further research.

7.2 Cross-linguistic categorization of languages according to PADDs

The results shown in Section 6.2 suggest that PADDs seem to categorize languages into several groups. These categorizations might reflect syntactic characteristics of individual languages, such as word order, yet it would be too simplistic to argue that longer dependency distances are the result of a particular word order; German and Hindi are languages of SOV word order, and it seems that they contain average dependency distances longer than other languages in the PUD, but this does not mean all the SOV languages to prefer longer dependency distances (e.g., Turkish). Besides, as far as the PADD is concerned, Chinese and Japanese are close to each other, yet their word orders are different. While the results in Section 6.2 do not strongly contradict our claim that we can obtain an objective, quantitative measure to indicate cross-linguistic variation of the syntactic-structural setting of human languages, we also need to extend our research into syntactic characteristics that can result in longer or shorter PADDs across languages, which motivates our investigations of TADDs.

7.3 Diversity of TADDs

The TADDs of the seven languages with respect to the above-mentioned five dependency types clearly show a certain level of diversity of dependency distances across different dependency types. For example, German and Hindi are distinct from other four languages in terms of their TADDs, indicating their preference for longer dependency distances between the verb of a clause and its subject noun, and between a noun and the verb of a relative clause modifying the noun.

8. Conclusion

This article explored the three types of dependency distances of the sentences in a multilingual parallel corpus, in order to verify the expectation that we can obtain an objective, quantitative measure to indicate cross-linguistic variation of the syntactic-structural setting of human languages. The analysis of PUD revealed that language-wise average dependency distances supported the claim that natural languages prefer shorter dependency distances, while pair-wise average dependency distances seem to categorize languages into several groups, and type-wise average dependency distances seem to provide us with fine-grained quantification of syntactic properties of individual natural languages.

Acknowledgement
This work was supported by JSPS KAKENHI Grant Number 20K00583.

References

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence for dependency length minimization in 37 languages. In *Proceedings of Natural Academy of Science*, 112(33):10336-10341. https://doi.org/10.1073/pnas.1502134112

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec P. Marantz, Yasushi Miyashita, and Wayne O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium (pp. 95-126)*. MIT Press, Massachusetts, US.

Daniel Gildea, and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286-310. https://doi.org/10.1111/j.1551-6709.2009.01073.x

Daniel Grodner, and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, 29(2):261-290. https://doi.org/10.1207/s15516709cog0000_7

Haitao Liu. 2007. Probability distribution of dependency distance. *Glottometrics*, 15:1-12.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159-191.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171-193. https://doi.org/10.1016/j.plrev.2017.03.002

Masanori Oya. 2013. Degree centralities, closeness centralities, and dependency distances of different genres of texts. In *Selected papers from the 17th Conference of Pan-pacific Association of Applied Linguistics*, 42-53.

Daniel Zeman. 2015. Slavic Languages in Universal Dependencies. In *Slovko 2015: Natural Language Processing, Corpus Linguistics, E-learning*. Slovakia.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, et al. 2020. *Universal Dependencies 2.7*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-3424.